热度分析技术在舆情吹哨系统中的应用

郑创伟 谢志成 陈少彬 邢谷涛 陈义飞
（深圳市创意智慧港科技有限责任公司，广东 深圳 518034）

摘要：【目的】为提高报业集团舆情相关工作的准确度和效率，文章研究热度分析技术在吹哨系统中的实际应用效果。【方法】提出热度及关联度计算，通过热度话题计算、关键词的关联相关度分析及关联热度计算，最后完成事件热度预测。【结果】通过热度分析技术实际应用，满足日常工作中的热点话题捕捉和及时跟踪，对舆情管理具有重要实施。【结论】通过本研究证明了吹哨系统中所使用的热度计算和关联热度计算等方法，极大地提高了吹哨系统的精确性，使用户可以从海量新闻信息中高效、智能地获得受关注、感兴趣、有价值的目标新闻信息，从而更加有力地支撑舆情监测、新闻追踪、新闻生产等业务工作。

关键词：舆情；吹哨系统；热度；关联热度；相关度 **中图分类号：**P413 **文献标识码：**A
文章编号：1671-0134（2023）05-134-05 **DOI：**10.19483/j.cnki.11-4653/n.2023.05.031
本文著录格式：郑创伟，谢志成，陈少彬，邢谷涛，陈义飞．热度分析技术在舆情吹哨系统中的应用 [J]. 中国传媒科技，2023（05）：134-138.

导语

舆情热点事件是当下互联网时代不可避免的事物之一，舆情热点事件发生后，民众往往会迅速对其热切关注，并且在此过程中民众会持续发表对该事件的观点、态度或表达一定的情绪。[1] 这类网络舆情热点事件从开始到发生一段时间后，最终往往会形成一个聚焦点，代表了网民的核心情绪和利益诉求。

在当前数据爆炸的时代，如何结合新闻信息的海量历史数据，为编辑、记者等新闻媒体从业者提供快速、精准、“千人千面”的个性化新闻线索推荐和智能吹哨预警支持，增强舆情态势感知能力和新闻洞察力，有效提升办公效率和新闻创造能力，是当前需要解决的问题。[2] 为解决这一问题，利用热度分析技术可以从海量新闻信息中高效、智能地获得受关注、感兴趣、有价值的目标新闻信息，从而更加有力地支撑舆情监测、新闻追踪、新闻生产等业务工作。

1. 热度分析技术相关研究

通过对热度分析相关文献整理，发现网络舆情热度分析可以从两个角度来进行。第一是从用户角度出发，分析用户在论坛、微博等平台上发布的话题情况，话题是由用户对事件进行描述所产生的，热点话题和普通话题的主要区别在于用户使用多少信息量来对其进行描述、消耗了多少网络资源，以及话题持续讨论的时间等。第二是从媒体角度出发，分析新浪、搜狐等新闻网站对热点事件转发、排名等情况。一个话题的出现与传播，是经过大众广泛讨论并且媒体进行报道和转载之后产生的，其中是否能成为热点话题，往往会根据报道数量及频率来进行衡量。[3]

近年来对网络舆情分析的研究已经逐步深入到了普通学者的实验课题探讨中。课题一般聚焦于在微博、微信、论坛等社交网络或应用中，这些社交场景中存在大量的活跃用户，一旦有热点话题出现，其传播速度会以指数级增长。热点网络舆情主要是依托网络进行传播，一个舆情事件被大众关注、评论、传播，从而引起更广泛的社会关注。在热度分析方面，国内研究者运用影响力传播模型描述热点事件，这种模型通过对关键词传播次数进行计数，数值大则代表影响力高，反之代表影响力较低。影响力传播模型可以用于评判社交网络中不同使用者之间所产生的交互程度。同时，通过分析话题的相关消息，以及转载次数等来评判其是否属于热点话题，利用用户关注度来构建影响力传播模型，通过关键词的传播次数反映某个事件影响力的大小。另外还有学者提出通过时间单元检测发现热点话题，即将某一话题限定在单元时间内，然后根据其特征分布情况来确定特征单元，再对其进行重组，最后生成热度话题，以及进一步确定出该热度话题所发生的时间段，达到更加精准预测的目的。[4]

本研究的热度分析技术主要是针对网络大众感兴趣的话题进行研究，使算力能聚焦于用户关注的话题，避免资源浪费。通过计算话题的热度，可以对不同话题的影响力进行排序，使得在吹哨系统中能够对排名靠前的话题进行预警。从而根据预警信息提前做出相应准备，尤其是当遇到极端情绪等，可以对其进行正确引导，避免话题对其他民众产生二次负面影响，成

为社会不稳定因素。针对不同话题影响力，吹哨系统还可以采取不同级别进行处理，更加精准地开展引导工作，提高舆情分析的有效性。

2. 热度及关联热度计算

2.1 热度计算

在本吹哨系统中，要实现从热点话题的发现及预测，两者对媒体行业都至关重要。而现有的研究成果大多使用的方法是进行热度计算，再结合以往经验数据来进行验证，判断其是否具有有效性。这种方式往往具有一定滞后性，无法在一个话题刚出现的时候就能有效预测其发展趋势，无法有效帮助政府部门及时、精准地调控舆论方向，也无法根据设定的监测规则来持续跟踪监测话题。因此，本研究采用Z算法对文章热度、敏感度等进行分析和归类，并将分析和归类结果保存，以便能够及时发现热点话题。[5] 具体过程如下。

首先，将语义分解后的新闻舆情数据，即词语化的数据，进行二元分布统计，统计各词语出现的次数，得出二元分布统计结果。

接着，将二元分布统计结果利用标准分数Z-Score算法进行计算，得到各词语的热点值。公式如下：

$$Z=\frac{X-\overline{X}}{S}$$

其中，公式中 X 为词项出现次数；$\overline{X}$ 为词项出现次数平均数；S 为标准差；结果 Z 是以标准差为单位的离均差，用以表示词语的热点值。[6]

将热点值大于预设的热点上限阈值的值存入热点词库中的热点活跃词库，将热点值小于预设的热点下限阈值的值存入热点词库中的热点惰性词库；热点词库与领域词库相关联，领域词库包括新闻、博客、论坛、社交网站等领域；每个热点词库中的热点词来源于哪些领域都可以进行对应查询。

再根据词语热点值和预设的热点词库判定词语化数据中的热点词的共现阈值。

根据新闻舆情数据中出现的词项，通过如下公式计算热点活跃词的共现阈值P1：

$$P1=\frac{W_x\cap W_h}{W_x}$$

其中 Wx 为新闻词项集合，Wh 为热点活跃词集合。再通过如下公式计算热点惰性词的共现阈值 $P2$：

$$P2=\frac{W_x\cap W_c}{W_x}$$

其中 Wx 为新闻词项集合，Wc 为热点惰性词集合。然后，根据热点活跃词和热点惰性词的共现阈值 $P1$ 和 $P2$，进行线性加权计算，得到热度值。热度值的计算公式如下：

$$H=\sum_{i=1}^{n}z_i(P1-P2)$$

其中 Zi 为第 i 个词语的热点值，$P1$ 为热点活跃词共现阈值，$P2$ 为热点惰性词共现阈值。然后，根据热度值对新闻舆情数据进行热度判定，对热点值根据预设的热度等级评判标准进行等级判定；将符合热度等级评判标准的新闻舆情数据归档至热点文档，将不符合热度等级评判标准的新闻舆情数据归档至非热点文档。[7]

在敏感度分析上，将热点活跃词库与预设的敏感词库进行比对得到热点活跃词库中包含的敏感词数量，再通过下述公式计算敏感值作为新闻敏感度S：

$$S=\frac{Ws}{Wn}$$

其中 Ws 为包含敏感词数量，Wn 为领域词库中新闻中的热点活跃词数量。

2.2 关联相关度分析

舆情预测就是需要对话题未来的趋势做出判断，一般来说相关话题的热度值越高则话题成为热点的概率也越大，也就是说所需要预测的话题成为热点的概率与其相关话题热度或数量成一定的关联关系。话题间的关联关系分析主要包含了对时间、地点、人物及行为等不同类型的词特征之间的关联度计算，以及对其进行加权。[8]

2.2.1 时间相关度计算

话题的时间相关度主要是指两个话题发生的时间差是否在一个指定的范围内。需要计算时间的间隔并以之判定相关度，如果在范围内，则认为两个话题在时间上是关联的，且时间间隔越短，则关联性越强，公示如下。其中，$time(T_1)$ 代表某一个话题的时间，T_i 和 T_j 则代表分别需要预测相关度的两个话题。如果需要分析话题出现的先后顺序，则将 $time(T_i)$ 按照时间顺序进行排列即可。

$$Rel^T(T_1,T_2)=\frac{time(T_1)-time(T_2)}{max_{Ti,Tj}Rel^T(T_i,T_j)}$$

2.2.2 地点相关度计算

在话题中的地点名称等信息是计算该相关度的主要依据，用主要地点间的距离来计算该相关度值。因此需要构造一个地点相关的名词集合，具体到城市的区级或农村的乡级，并且要对应更高行政区域建立一个层次树。如果预测的话题所属地域之间，距离在一定的范围内，则可以认为其是相互关联的，关联强度则可以根据间隔距离计算，距离越近则说明关联程度越高。公式如下，其中 $locate(T_1)$ 表示话题发生的主要地点，其与 $locate(T_2)$ 之差则表示两个话题发生地点在层次树上的路径长度。

$$Rel^L(T_1,T_2)=\frac{locate(T_1)-locate(T_2)}{max_{Ti,Tj}Rel^L(T_i,T_j)}$$

2.2.3 人物相关度计算

人物相关度主要是指被预测话题所涉及的人物或机构是否相互之间关注或有其他关系，如果存在好友或其他关系，则认为这两个话题在人物上是关联的。但往往在实际应用中，微博或微信好友关系是无法取得的，因此可以利用话题中的人名进行计算，例如通过人名重复的数量来进行计算。公式如下，其中 *people*（T_1）为某一话题中涉及人物名称等的集合，T_i 和 T_j 则代表两个需要预测的话题。

$$Rel^P(T_1,T_2)=\frac{people(T_1)\cap people(T_2)}{max_{Ti,Tj}Rel^P(T_i,T_j)}$$

2.2.4 行为相关度计算

行为相关度主要是收集话题行为的特征词来进行计算，如果涉及的行为相同或相近，则认为其是相关的。公式如下，其中 A_1 和 A_2 代表两个话题中行为特征词的集合，*maxsim*（*w*，A_i）则为词语语义的相似度，*IDF*（*w*）是根据预料库中词信息量统计得到。

$$Rel^A(A_1,A_2)=\frac{1}{2}\left(\frac{\sum_{w\in A_1}(maxSim(w,A_2)*IDF(w))}{\sum_{w\in A_1}IDF(w)}+\frac{\sum_{w\in A_2}(maxSim(w,A_1)*IDF(w))}{\sum_{w\in A_2}IDF(w)}\right)$$

2.3 关联热度计算

针对舆情热度的计算与预测研究当前在学术界已经取得了一定的成果，但大部分算法主要是针对数据进行分析，没有对网络舆情本身的特点进行数据分析，尤其是忽视了网络信息之间的互联性。因此本研究在基于热度计算的基础上，结合了关联分析的思想，综合考虑时间、地点、人物、行为的相关性，对不同属性的相关关系进行挖掘，构建具有关联关系的舆情热度预测模型，通过分析相关事件或信息的关系，对热度建立相应的回归模型，使得热度值更加贴近实际情况。

关联热度计算主要就是根据话题热度按时间对其进行分片，再根据命名实体对其进行识别，例如通过时间信息计算出时间相关度、通过地点信息计算出地点相关度、通过人物信息计算出人物相关度、通过行为数据计算出行为相关度，最后建立相关关系连接图。[9]

在本吹哨系统中，建立新闻话题间的关系图，再计算出热度值，并将其设置为初始权重值，用于某一时间段内的关联热度计算。热度计算完成后，再利用相关度算法来对话题热度的变化趋势进行预测和分析，实现吹哨系统预警。

2.3.1 建立话题间关系

设定 A=<V，E> 为v_1的关系图，如图 1 所示，其中v_1为给定话题，集合$\{v_1,v_2,...,v_{n-1},v_n\}$为检索到的与$v_1$相关的话题集合，$E=\{v_1v_2,v_1v_3,...,v_{n-1}v_n\}$是边的集合，值为话题间的相关程度，当且仅当两个顶点v_1v_2间关联度不小于阈值时，边v_1v_2存在。

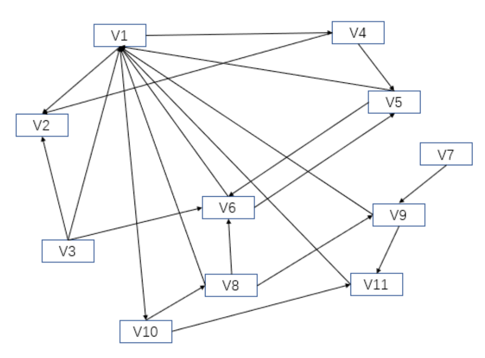


图 1 话题关系联系图

建立关系连接图后，下一步将图转化为矩阵形式，矩阵中的行、列表示联系图中的点，矩阵中的值表示关系图中点间的度。如以下图 2 所示，其中 R_{ij} 是节点 *i* 和节点 *j* 间的相关程度，相关度小于阈值的即不存在边 *ij* 则值为 0。

$$\begin{bmatrix} & v(1) & v(2) & \cdots & v(i) & \cdots & v(j) & \cdots & v(n)\\ v(1) & R_{11} & R_{12} & \cdots & R_{1i} & \cdots & R_{1j} & \cdots & R_{1n}\\ v(2) & R_{21} & R_{22} & \cdots & R_{2i} & \cdots & R_{2j} & \cdots & R_{2n}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ v(i) & R_{i1} & R_{i2} & \cdots & R_{ii} & \cdots & R_{ij} & \cdots & R_{in}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ v(j) & R_{j1} & R_{j2} & \cdots & R_{ji} & \cdots & R_{jj} & \cdots & R_{jn}\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ v(n) & R_{n1} & R_{n2} & \cdots & R_{ni} & \cdots & R_{nj} & \cdots & R_{nn} \end{bmatrix}$$

图 2 矩阵形式列表图

2.3.2 相关话题关联重要度计算

定义变换矩阵 M，公式如下：

$$M=d\cdot R+(1-d)$$

其中，d 为阻尼系数，范围介于 0 至 1 之间。该矩阵主要作用在于衡量每个点对待预测点的影响力。矩阵 M 具有唯一稳定分布$h=M^T\cdot h$。该模型的矩阵表示为：

$$h=[dR+(1-d)]^T\cdot h$$

得到的 h 值则可以用于表示话题在关系图中的重要程度。

2.3.3 热度预测

在吹哨系统中，需要对具有少量当前信息的舆情短期热度趋势进行预测，判断该话题是否会成为热点话题，本研究采用灰度预测方法来进行趋势预测。通常使用 GM（1，1）模型来对话题热度进行预测，计算过程如下 [10]：

a. 输入初始序列 $X^{(0)}=(x^{(0)}(1),x^{(0)}(2),K,x^{(0)}(n))$；

b. 对初始序列进行一次累加生成，

$$X^{(1)}=\left(x^{(1)}(1),x^{(1)}(2),K,x^{(1)}(n)\right)X1;$$

c. 生成 X1 的紧邻均值序列

$$Z^{(1)}=(z^{(1)}(2),z^{(1)}(3),K,z^{(1)}(n))$$

$$z^{(1)}(k)=0.5x^{(1)}(k)+0.5x^{(1)}(k-1)$$

d. 即 GM（1，1）的灰微分方程模型为

$$x^{0}(k)+az^{(1)}(k)=b$$

式中 a 为发展系数，b 为灰色作用量。设 $\widehat{\alpha}$ 为待估参数向量，即 $\widehat{\alpha}=(a,b)^{T}$，则灰微分方程的最小二乘估计参数列满足

$$\widehat{\alpha}=(B^{T}B)^{-1}B^{t}Y_{n}$$

其中，$B=\begin{bmatrix}-Z^{(1)}(2) & 1\\ -Z^{(1)}(3) & 1\\ \cdots & \cdots\\ -Z^{(1)}(n) & 1\end{bmatrix}$，$Y_{n}=\begin{bmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ K\\ x^{(0)}(n)\end{bmatrix}$

e. 求得微分方程得解为

$$\hat{x}^{(1)}(k+1)=\left[x^{(t)}(0)-\frac{b}{a}\right]e^{-ak}+\frac{b}{a}$$

f. 还原到原始数据，得到

$$\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)$$

得到热度趋势预测区间，结束。

2.4 热度预测

在笔者实际工作中主要用到的方法是基于事件关联的方法对舆情趋势进行预测，并判断其是否成为热点话题。这种模型主要是基于假设“事件是相互关联且相互影响的”，事件与事件之间存在着一定的联系，并且可能会相互影响或约束，其算法框架如图 3 所示 [11]：

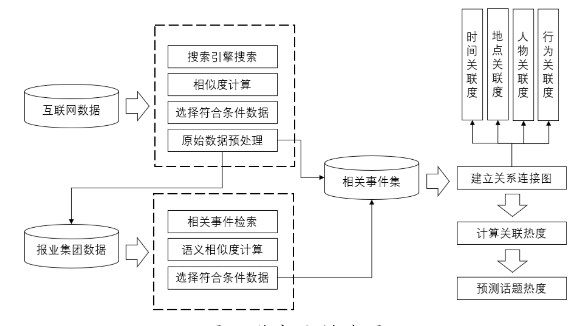


图 3　热度预测框架图

能够看出其具体流程主要包括 [12]：

（1）检索出近段时间内与待预测话题相关的事件，在设置检索词时需注意特征词的选取。

（2）对集团本地数据库进行检索，与互联网上的检索进行比对，并分析话题间的相互关系，获得与舆情事件有关的文字信息数据。但在数据收集后需要对数据信息进行去噪等处理，保证一定的准确性。

（3）对整理出的文本信息采用聚类算法分析，提取出其可能包含的话题数量。

（4）对文本数据进行时间排序，按照实际需求来设定时间段，在每一个时间段根据事件发生的时间、人物、地点、行为等，计算出话题间的相关度，从而得到所有话题的关系，即关系连接图。

（5）分析不同话题的重要程度，并且预测关联热度，最终计算出该话题或信息成为热点的可能性。

3. 实验结果及分析

3.1 实验设计

本吹哨系统在对舆情热度进行预测后，进一步利用后验差检验方法来验证实验效果，具体步骤包括：

（1）计算原始序列的平均值；

（2）计算原始序列的均方差 S1；

（3）计算残差均值；

（4）计算残差均方差 S2；

（5）计算 S2 与 S1 的比值 C

（6）计算小残差概率 P

3.2 实验结果

本研究分别使用 P 值和 C 值来衡量突发舆情的预测效果，并设计了相应的后验差检验判别参照表（见表 1）。

表 1　后验差检验判别参照表

P	C	模型精度
>0.95	<0.35	优
>0.80	<0.5	合格
>0.70	<0.65	勉强合格
<0.70	>0.65	不合格

在数据库中对“孙小果案”相关数据进行热度预测，分别包括长期预测、短期预测、普通灰度预测和关联热度预测，所得到的实验结果如下：

表 2　实验结果表

	P	C
长期预测	0.7692	0.6038
短期预测	0.8385	0.3846
普通灰度预测	0.9125	0.4129
关联热度预测	1	0.0192

从表 2 结果看出，关联热度计算的方法对突发舆情的预测效果非常好，验证了该吹哨系统中所使用的热度分析技术的可行性和有效性。

结语

本研究对报业集团吹哨系统中所使用的热度计算、关联相关度分析、关联热度计算，以及热度预测等进行了深入分析，分别列出了相关公式和模型中涉及的相关因素，例如时间、地点、人物及行为等不同类型的词特征，从而计算出事件之间的关联度，并预测是否会发展成为热点事件。通过上述方法和实际应用，

证明报业集团吹哨系统具有较好的精确性，使用户可以从海量新闻信息中高效、智能地获得受关注、感兴趣、有价值的目标新闻信息，从而更加有力地支撑舆情监测、新闻追踪、新闻生产等业务工作。政府也可以借助该系统引导舆情方向，对重大舆论事件可以快速做出反应。这可以在一定程度上抑制大众对舆论事件产生的消极情绪，将有利于政府正确引导舆情发展趋势，以及保持社会和谐稳定。

参考文献

[1] 梁修明 . 新媒体环境下公共危机传播治理路径 [J]. 中国传媒科技，2019（5）：48-50.

[2] 冯小东，李卓雅，史志慧 . 基于网络舆情热度的自然灾害影响评估分析 [J]. 情报探索，2020（1）：16-22.

[3] 袁然 . 全媒体传播中数据技术的应用实践 [J]. 中国传媒科技，2021（7）：21-23.

[4] 高萍，周恩 . 大数据背景下政府危机公关的舆情引导及对策研究——以政务微博为例 [J]. 阴山学刊，2019（6）：88-94.

[5] 毛通，谢朝德 . 基于百度大数据的信用舆情指数构建与实证研究 [J]. 征信，2020（1）：11-20.

[6] 王文妤，阴雪颖 . 基于模糊评价法政府实时监控网络舆情热度模型构建 [J]. 中国管理信息化，2019（23）：170-173.

[7] 邹佳成，马远远，刘婷，唐伯超，刘振国，高辉 . 基于大数据的酒业舆情信息监测平台 [J]. 酿酒科技，2020（3）：129-135.

[8] 张丕翠，杨建武，施水才 . 网络空间的舆情态势感知 [J]. 信息安全研究，2019（11）：1013-1020.

[9] 李靖云 . 新媒体环境中热点事件的舆情治理策略 [J]. 新闻潮，2019（10）：44-45+48.

[10] 曾润喜 . 网络舆情治理的关键是“治未病”[J]. 中国传媒科技，2018（12）：12-14.

[11] 张源淇 . 影响网络舆情热度评价的识别因素探讨 [J]. 新闻研究导刊，2022（1）：127-129.

[12] 王茜仪，杜明坤，张山 . 基于深度学习的网络舆情热度研究 [J]. 无线互联科技，2020（22）：16-17.

作者简介：郑创伟（1978-），男，广东汕头，高级工程师，研究方向为大数据、人工智能；谢志成（1980-），男，广东汕头，中级职称，研究方向为大数据、云计算；陈少彬（1973-），男，广东揭阳，中级职称，研究方向为大数据；邢谷涛（1984-），男，海南文昌，研究方向为云计算；陈义飞（1981-），男，广东湛江，中级职称，研究方向为大数据。

（责任编辑：张晓婧）

（上接第120页）

[3] 邹洁 . 医学类数字教材的出版流程——以《医学免疫学》数字教材为例 [J]. 出版广角，2015（12）：165-167.

[4] 钱新艳 . “互联网 +”时代教材出版中的——以医学类高等教育教材为例二维码应用及“纸”“数”融合实践思考 [J]. 科技传播，2020（10）：70-73.

[5] 杨永洁，王晶，严瑛，等 . 新型冠状病毒肺炎疫情防控期间“停课不停学”线上教学状况的调查与分析 [J]. 青岛大学学报（医学版），2020（5）：601-604.

[6] 鲍莹 . “新”设计 “新”结构 “新”教学——人教版高中地理教材必修第一册解读 [J]. 课程教材教学研究（中教研究），2022（Z1）：6-8.

[7] 陈思铧 . 现代医学教育模式的三次转变及影响因素研究 [D]. 石家庄：河北医科大学，2022.

[8] 高原，王莹，王哲，等 . 翻转课堂教学模式在病理学的实践 [J]. 中国中医药现代远程教育，2022（23）：19-20.

[9] 戴俊程，王辉，方蕊，等 . 以思维导图为基础的互动式教学应用于《医学研究的数据管理与分析》课程的思考 [J]. 课程教育研究，2018（3）：221-222.

[10] 荣康，李祥林，殷志杰，等 . VR 技术在医学影像专业教学中的应用探索与思考 [J]. 中国继续医学教育，2022（23）：184-188.

[11] 李娟 . 元宇宙背景下 AR 与 VR 技术在新媒体的应用及影响探析 [J]. 中国传媒科技，2022（12）：157-160.

[12] 宋永刚，张会 . 从实践角度试论纸数联合出版在出版社数字化转型中的地位与价值 [J]. 科技与出版，2019（6）：75-78.

[13] 侯良健 . 融合出版中教材数字资源建设实践与思考 [J]. 中国传媒科技，2022（3）：129-131.

作者简介：雷媛（1986-），女，山西运城，人民卫生电子音像出版社，医学数字编辑，出版中级，研究方向为医学融合教材及数字教材建设。

（责任编辑：张晓婧）